Assignment 3: Wine Quality

Akanksha Kushwaha

02/09/2024

Overview

- Many wine brands are seeking new ways to maximize the success of their wines. Before making any decisions, it might be helpful to know which features contribute to a wine's quality. Knowing these features can enable a brand to make more intelligent decisions when making it. But what exactly are these features? Using ML techniques with wine data retrieved from the following website, I plan to answer this question. https://archive.ics.uci.edu/ml/datasets/wine+quality

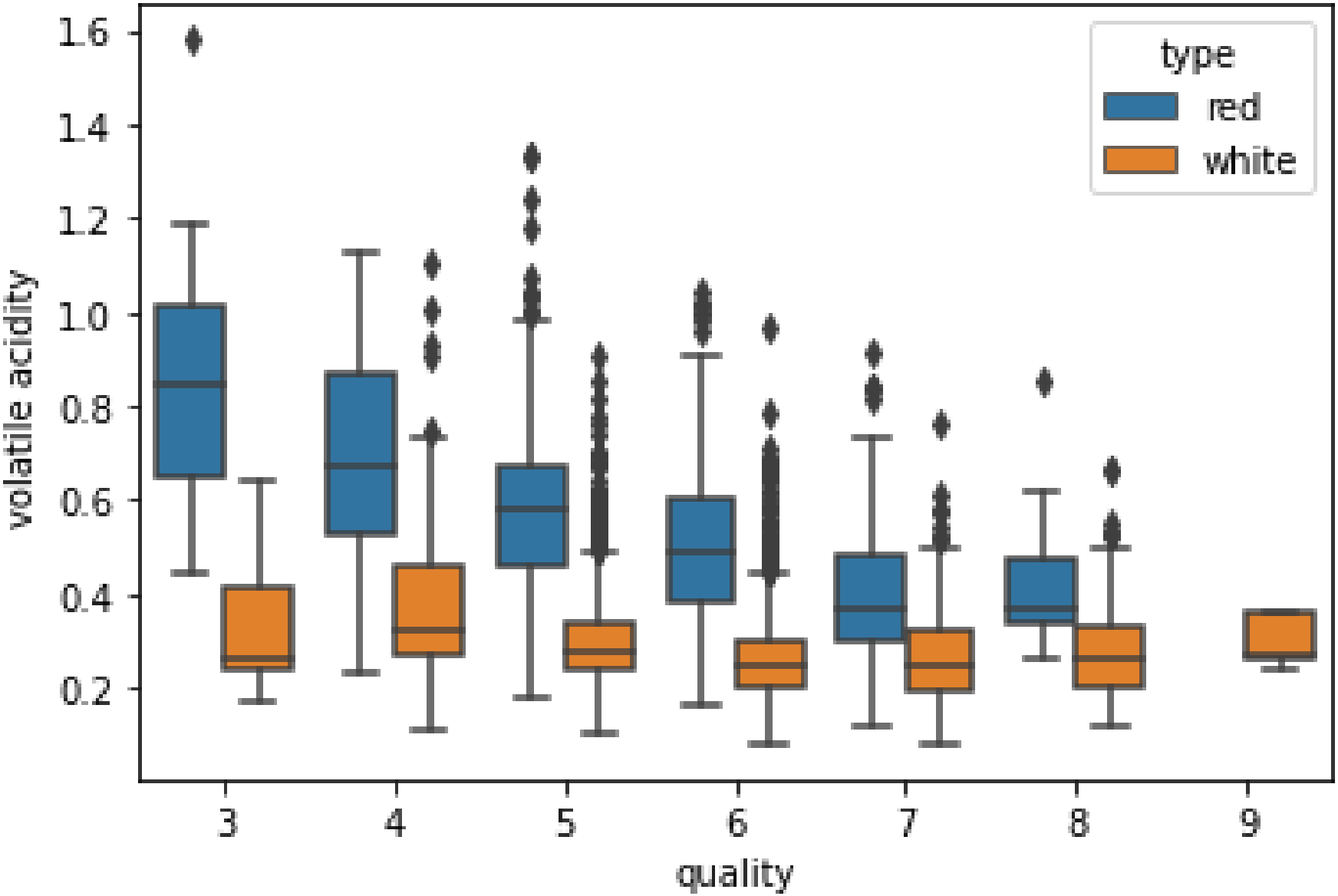
Data Wrangling

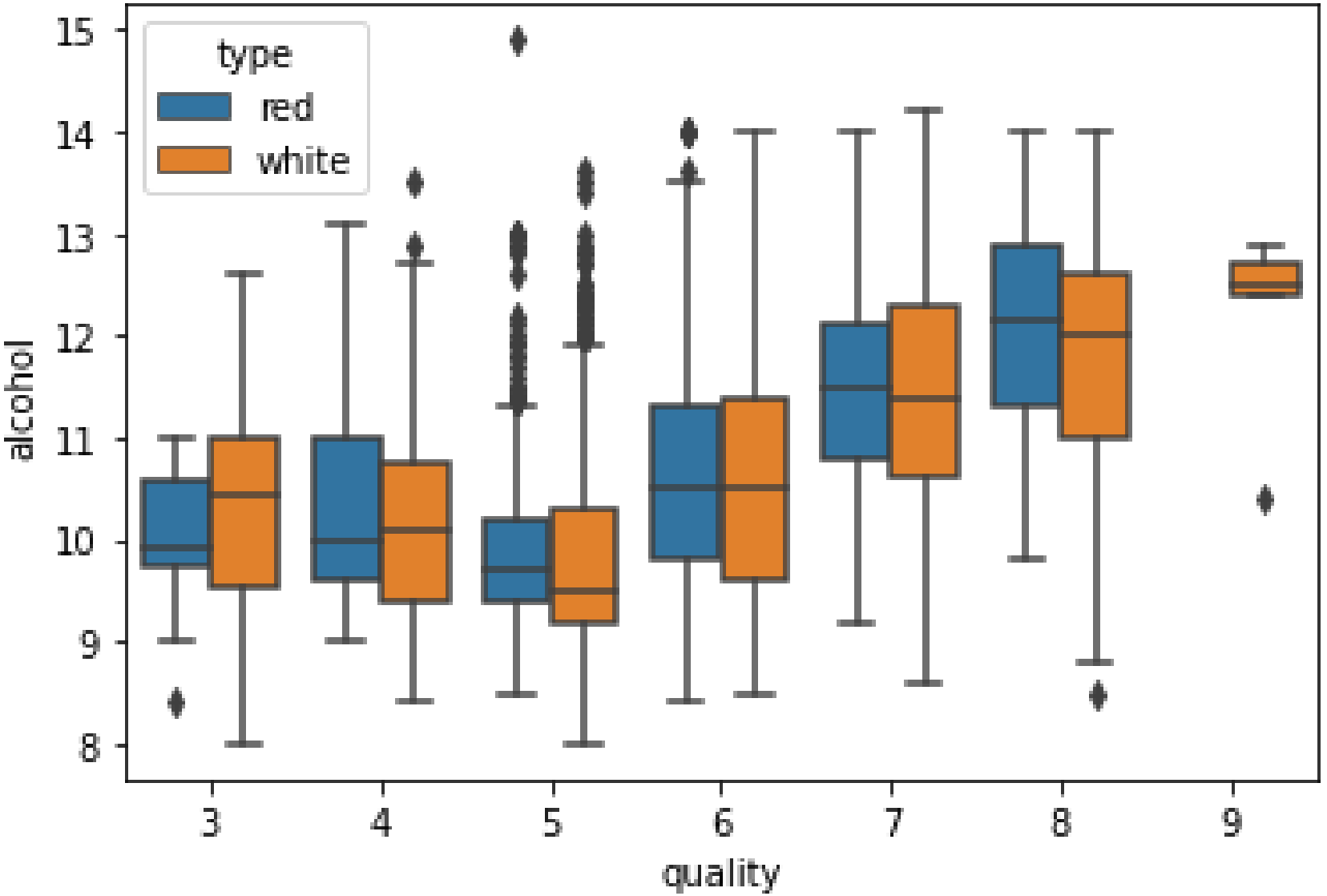
- The dataset for this project was wrangled by another party prior to beginning this project.
- Using red and white wine samples, inputs include objective tests (PH values) and the output is based on sensory data (wine tasting by experts). Using a median of at least 3 evaluations, each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach to determine wine quality.

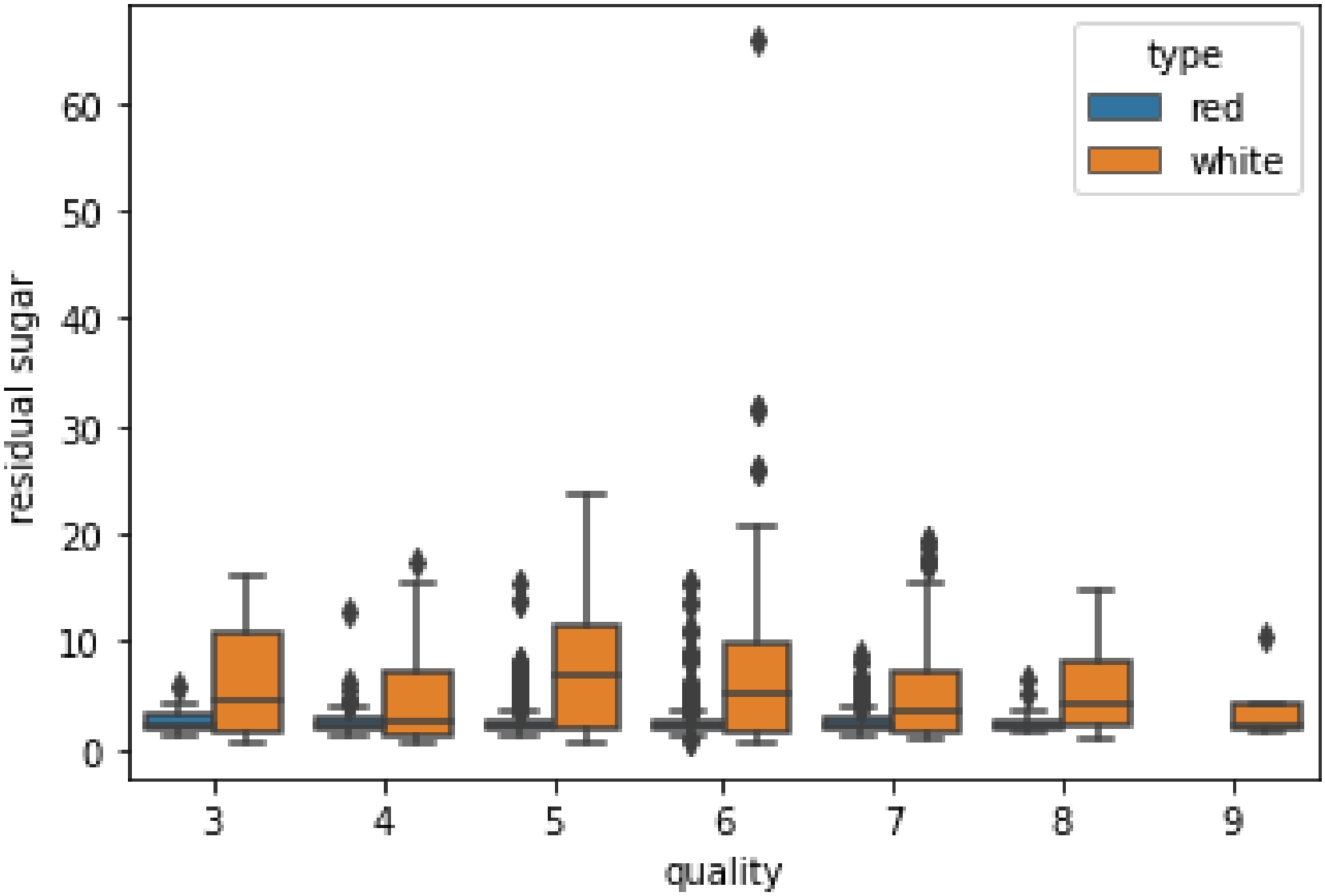
Exploratory Data Analysis

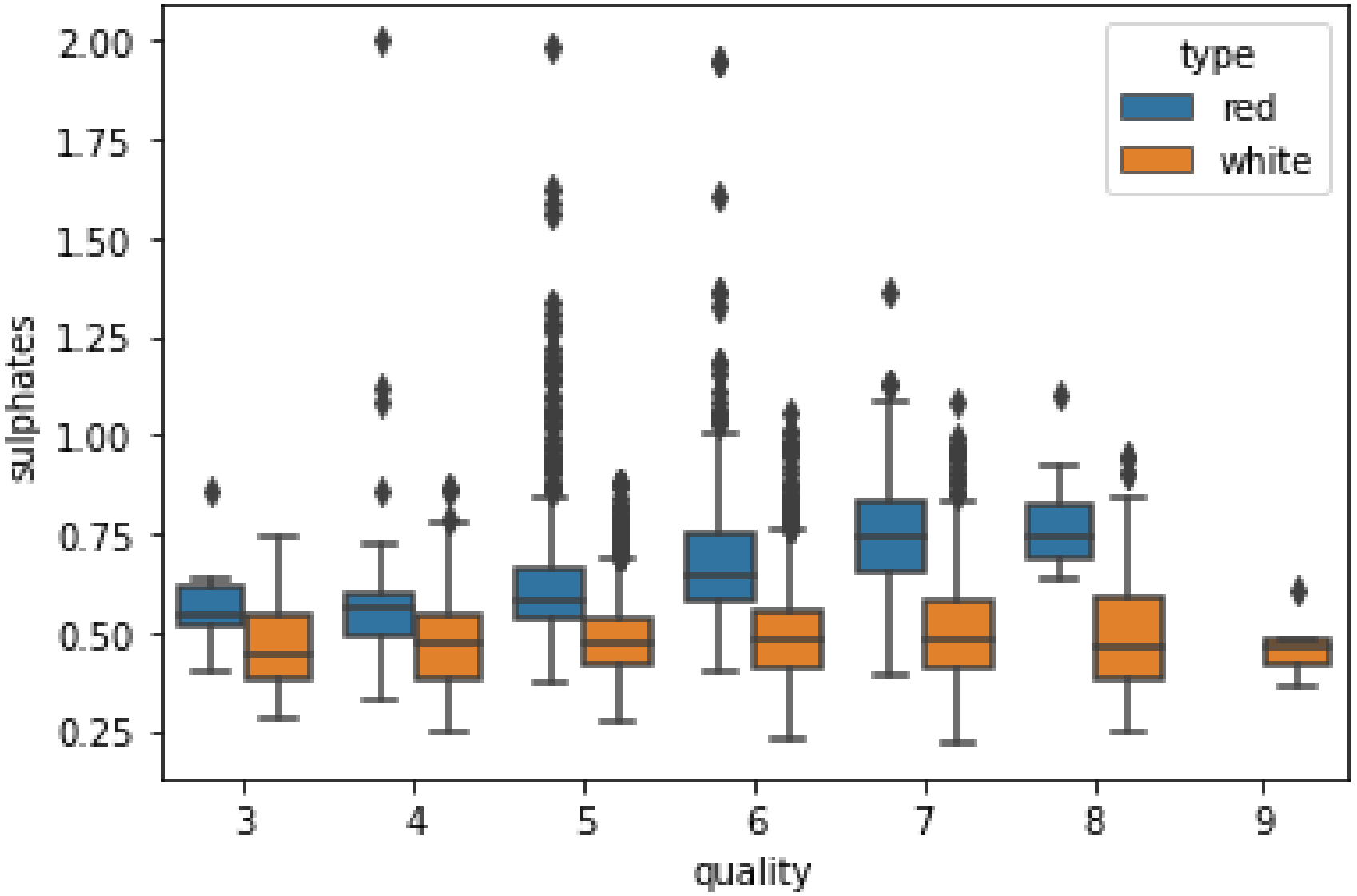
- Exploratory data analysis can be used to derive relationships between the wine quality and the various features available from the wine's profile and suggest improvements to the profiles that would increase the wine's quality. The quality feature of this set was the result of wine tasters opinions, ranging from 3 to 9, with the higher numbers being higher quality. The spread is shown below:

"quality" value counts:	
6	2836
5	2138
7	1079
4	216
8	193
3	30
9	5

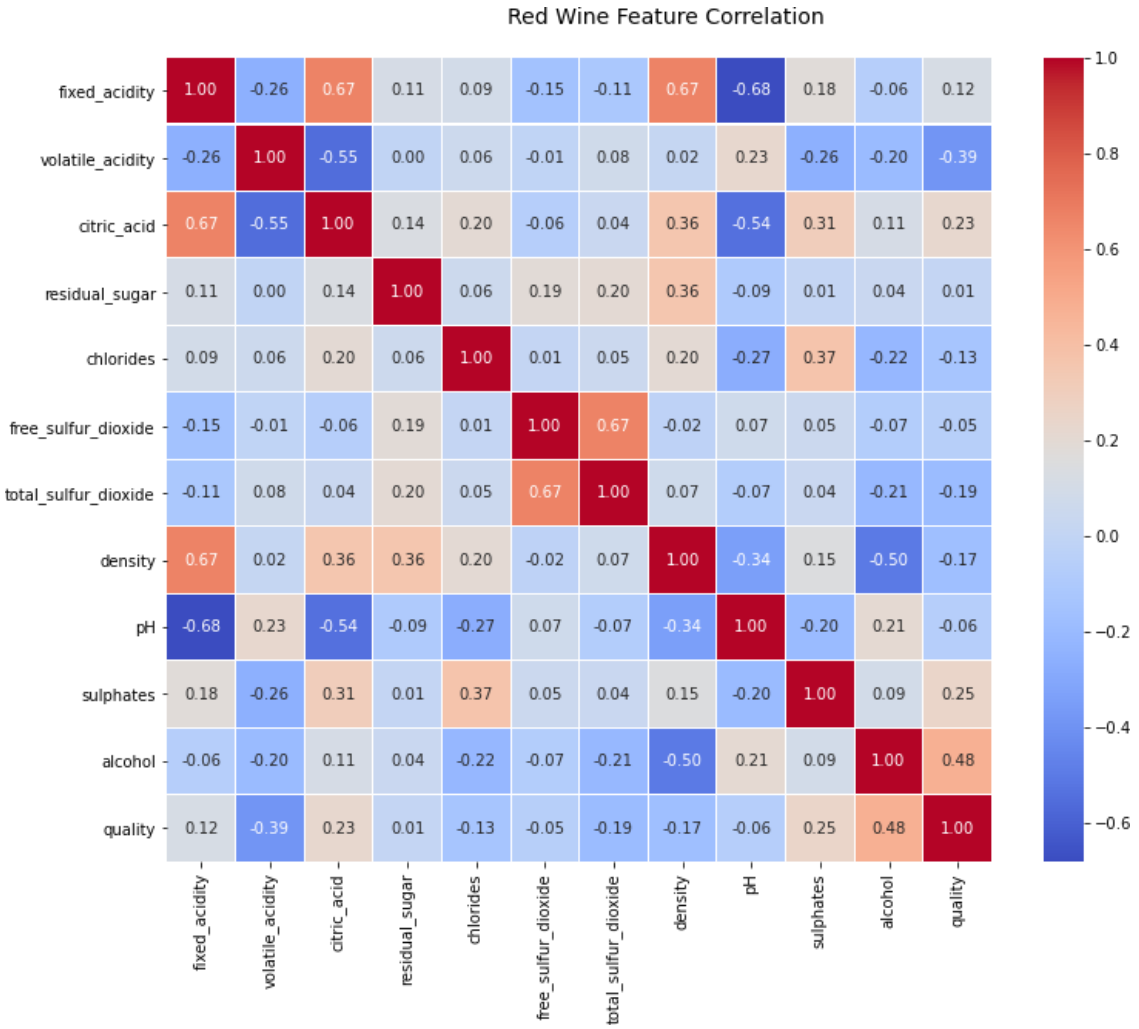

type
red
white
volatile acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
3
4
5
6
7
8
9
quality


type
red
white
alcohol
15
14
13
12
11
10
9
8
3
4
5
6
7
8
9
quality

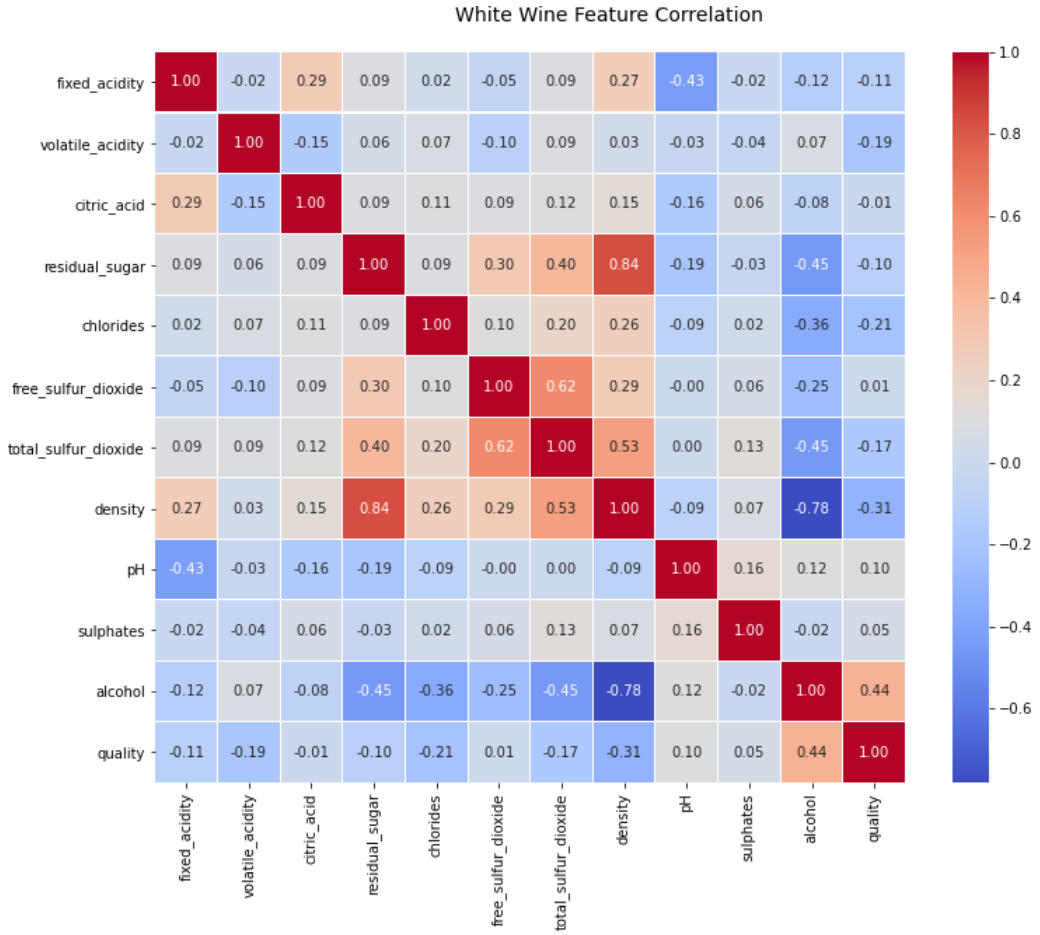

type
red
white
residual sugar
0
10
20
30
40
50
60
3
4
5
6
7
8
9
quality


type
red
white
sulphates
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
3
4
5
6
7
8
9
quality

The Red Wine Correlation Heatmap
fixed acidity has correlations ith citric acid, density, and pH.
free & total sulfur dioxide have a fairly high positive correlation



The White Wine Correlation Heatmap
alcohol and density have a fairly high negative correlation
free & total sulfur dioxide have a fairly high positive correlation



Preprocessing and Modeling

- Split the data into high and low quality wines, as well as red and white, for a total of two subsets of data: Red and White.
- The data was then scaled.
- I didn’t leave out any features due to high correlation.

Feature Importance

Red Wine

- Alcohol
- Volatile Acidity
- Sulphates

White Wine

- Alcohol
- Volatile Acidity
- Residual Sugar

Model Comparison

- LogisticRegression
- RandomForestClassifier
- KNeighborsClassifier

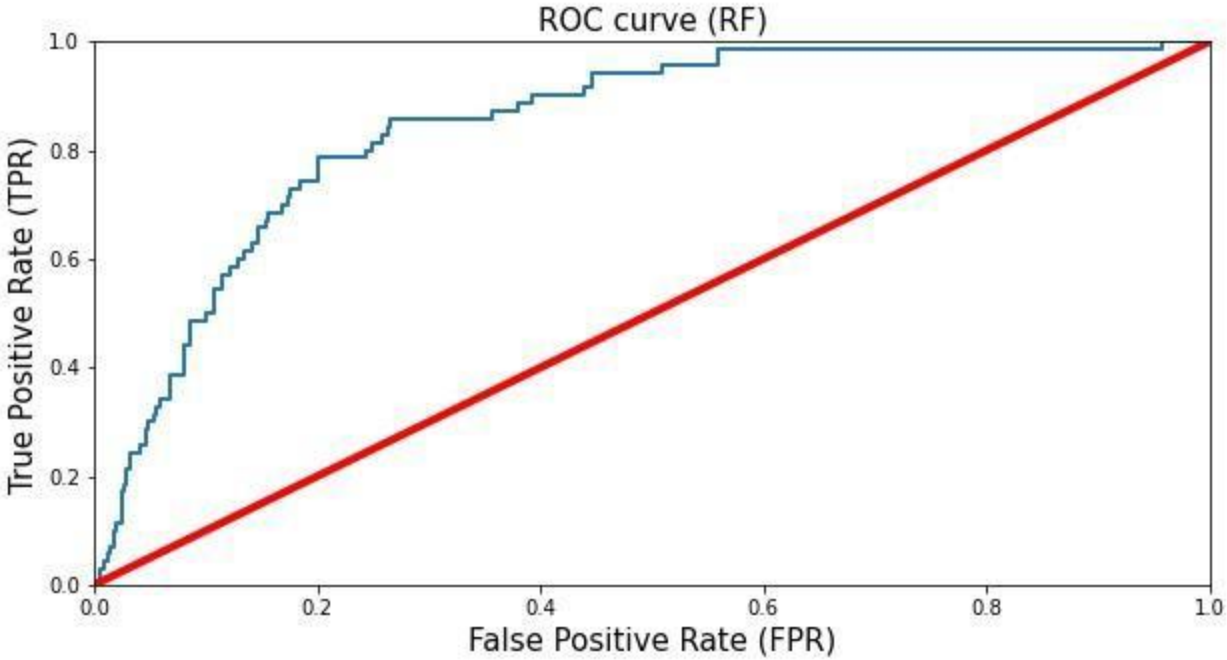
Red: 0.84 **White: 0.91**	
Red: 0.79 White: 0.75	
Red: 0.38 White: 0.77	

GridsearchCV Hyperparameter Tuning

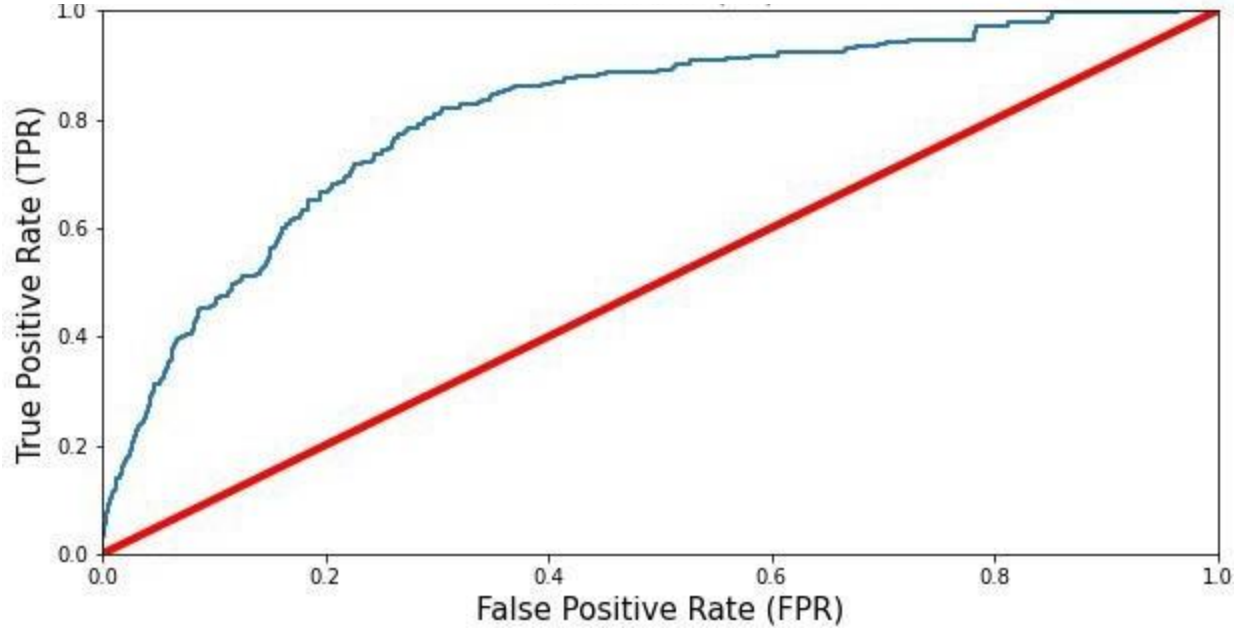
- Logistic Regression
- C=100, penalty='l1', solver='liblinear‘

- RandomForest Classifier
- max_depth=4, n_estimators=80

- Kneighbors Classifier
- algorithm='ball_tree', n_neighbors=100, weights='distance'


ROC curve (RF)
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0

Red


True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0

White

Classification Reports

Red:

	precision	recall	f1-score	support
Lo quality	0.88	0.99	0.93	410
High quality	0.76	0.19	0.30	70
accuracy			0.87	480
macro avg	0.82	0.59	0.61	480
weighted avg	0.86	0.87	0.84	480

White:

	precision	recall	f1-score	support
Lo quality	0.82	0.99	0.89	1166
High quality	0.72	0.14	0.24	304
accuracy			0.81	1470
macro avg	0.77	0.57	0.57	1470
weighted avg	0.80	0.81	0.76	1470

Conclusion

By analyzing the physicochemical data of red and white wines, I was able to create a model that can help industry producers, distributors, and sellers predict the quality of red wine products and have a better understanding of each critical feature. I found the Logistic Regression model performed better than the other two models. I determined three features most influential for both red and white wines. Red: volatile acidity, sulphates, and alcohol content. White: volatile acidity, residual sugars, and alcohol content. To be more specific, high-quality red wines seem to have lower volatile acidity, higher alcohol, and medium-to-high sulphates. Meanwhile, higher quality white wines also have low volatile acidity and high alcohol content, but differ in due to lower residual sugars.

This analysis comes with some limitations. First, the data set is unbalanced. A majority of the quality values were 5 and 6, which makes no significant contribution to finding an optimal model. These values make it harder to identify each features exact influence on a “high” or “low” quality of the wine, which was the main focus of this analysis. In order to improve the predictive model, more balanced data is needed. Another limitation worth mentioning is that the dataset only has 12 attributes, which reduces the accuracy of the predictive models. The solution for this is to include more relevant data features, such as the year of harvest, amount of brew time, or grape type. Different performance measures and/or machine learning techniques could also be utilized to find better performance and model comparisons.